Detecting Pretraining Data from Large Language Models

Year: 2023 **|** *Citations: 290* **|** *Authors: Weijia Shi, Anirudh Ajith, Mengzhou Xia*

Abstract

Although large language models (LLMs) are widely deployed, the data used to train them is rarely disclosed. Given the incredible scale of this data, up to trillions of tokens, it is all but certain that it includes potentially problematic text such as copyrighted materials, personally identifiable information, and test data for widely reported reference benchmarks. However, we currently have no way to know which data of these types is included or in what proportions. In this paper, we study the pretraining data detection problem: given a piece of text and black-box access to an LLM without knowing the pretraining data, can we determine if the model was trained on the provided text? To facilitate this study, we introduce a dynamic benchmark WIKIMIA that uses data created before and after model training to support gold truth detection. We also introduce a new detection method Min-K% Prob based on a simple hypothesis: an unseen example is likely to contain a few outlier words with low probabilities under the LLM, while a seen example is less likely to have words with such low probabilities. Min-K% Prob can be applied without any knowledge about the pretraining corpus or any additional training, departing from previous detection methods that require training a reference model on data that is similar to the pretraining data. Moreover, our experiments demonstrate that Min-K% Prob achieves a 7.4% improvement on WIKIMIA over these previous methods. We apply Min-K% Prob to two real-world scenarios, copyrighted book detection, and contaminated downstream example detection, and find it a consistently effective solution.